PCovR: An R Package for Principal Covariates Regression

Vervloet, Marlies; Kiers, Henk A L; Van den Noordgate, Wim; Ceulemans, Eva

Published in:
Journal of Statistical Software

DOI:
10.18637/jss.v065.i08

Publication date:
2015

Journal of Statistical Software

May 2015, Volume 65, Issue 8. *http://www.jstatsoft.org/*

PCovR: An R Package for Principal Covariates Regression

Marlies Vervloet
KU Leuven

Henk A. L. Kiers
University of Groningen

Wim Van den Noortgate
KU Leuven

Eva Ceulemans
KU Leuven

Abstract

In this article, we present **PCovR**, an R package for performing principal covariates regression (PCovR; De Jong and Kiers 1992). PCovR was developed for analyzing regression data with many and/or highly collinear predictor variables. The method simultaneously reduces the predictor variables to a limited number of components and regresses the criterion variables on these components. The flexibility, interpretational advantages, and computational simplicity of PCovR make the method stand out between many other regression methods. The **PCovR** package offers data preprocessing options, new model selection procedures, and several component rotation strategies, some of which were not available in R up till now. The use and usefulness of the package is illustrated with a real dataset, called `psychiatrists`.

Keywords: principal covariates regression, dimension reduction, multicollinearity, regression models, R.

1. Introduction

Principal covariates regression (PCovR) was proposed by De Jong and Kiers (1992) to deal with the interpretational and technical problems that are often encountered when applying linear regression analysis using a relatively high number of predictor variables – say, higher than 10. Indeed, when interpreting a particular regression weight, in principle all other predictors and corresponding regression weights have to be taken into account. Furthermore, chances get higher that at least some of the predictor variables will be highly correlated with a linear combination of the other predictor variables. In the latter case, parameter estimates

may become unstable, in that removing or adding one single observation can dramatically alter the regression weights, which is the so-called bouncing beta problem (Kiers and Smilde 2007).

In PCovR, the predictor variables are reduced to a limited number of components and the criterion variables are regressed on these components. Specifically, the components are linear combinations of the predictor variables that are constructed in such a way that they summarize the predictor variables as good as possible, but at the same time allow for an optimal prediction of the criterion variables. As the user may choose the extent to which both aspects (good summary of predictors, optimal prediction of criteria) play a role when constructing the components, PCovR is a flexible approach that subsumes principal components regression (Jolliffe 1982) and reduced-rank regression (Izenman 1975) as special cases. As is often the case with component techniques, the components have rotational freedom (including reflectional and permutational freedom) which can be exploited to enhance the interpretability of the PCovR parameters. Another attractive feature of PCovR is that a closed form solution exists, as optimal model estimates can be obtained by conducting one single eigenvalue decomposition. However, the flexibility of PCovR (number of components to be used, extent to which summarizing the predictors and predicting the criteria are emphasized, rotational freedom) has as downside that the user has to choose among a huge number of possible solutions. Up to now, no software was available to assist users in this daunting task.

To be sure, other dimension reduction methods exist for solving the above described problems, such as the R (R Core Team 2015) package **pls** (Mevik, Wehrens, and Liland 2013) for partial least squares (PLS; Wold, Ruhe, Wold, and Dunn III 1984), which is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=pls`. However, whereas PLS focuses explicitly on the prediction of the criterion block, PCovR allows to flexibly balance appropriate reduction of the predictors and accurate prediction of the criteria. Hence, it is no surprise that Kiers and Smilde (2007), who compared the performance of PCovR (using three different weighting schemes) and PLS in five different simulation settings, demonstrated that, for each setting, at least one PCovR weighting scheme yielded better or equally good results than PLS. The method that resembles PCovR the most is exploratory structural equation modeling (ESEM; Asparouhov and Muthén 2009), implemented in the commercial software package **Mplus** (Muthén and Muthén 1998–2011). However, ESEM does not have the unique combination of flexibility and computational simplicity that is typical for PCovR. In this paper we present a new R package, the **PCovR** package (Vervloet, Kiers, and Ceulemans 2015), available from CRAN at `http://CRAN.R-project.org/package=PCovR`, to perform principal covariates regression in R. Several rotation options are provided in this package, including some rotation strategies that were not available in R up to now, as well as some new model selection procedures.

The remainder of this paper is structured as follows: First, we will recapitulate PCovR analysis, by discussing the data and the associated preprocessing, model formulae, loss function, model estimation, and model selection. Next, we will describe the usage of the **PCovR** package, by giving a step-by-step overview of the available options.

2. PCovR analysis

2.1. Data

To run a PCovR analysis, two matrices are needed. A first matrix, $\mathbf{X}$, contains the information regarding the J predictors under study and the second, $\mathbf{Y}$, holds the data on the K criteria. These predictors and criteria are measured for the same N observations.

When applying dimension reduction methods, appropriate preprocessing of the data is important, as it will influence the obtained results. Here we consider two different forms of preprocessing: centering and scaling. As PCovR is based on the principles of principal component analysis, centering of $\mathbf{X}$ (i.e., setting the mean of each predictor to zero) is necessary to model the correlation or covariance structure of the data. Centering of $\mathbf{Y}$ is not necessary, but may enhance the interpretation of the regression weights and discards the need for an intercept (which can however be easily computed if desirable).

Regarding scaling, in component analysis, variables that have a larger variance influence the obtained components more. If such differences in variance may be arbitrary, e.g., caused by response tendencies or differences in response scales, it is advised to normalize the data (i.e., scale each variable to a variance of one). An additional advantage of normalizing the data, is that the obtained regression weights for a specific criterion variable can be compared in size.

2.2. Model

In PCovR, the J predictors are reduced to R new variables, called components:

$$\mathbf{X} = \mathbf{T}\mathbf{P}_\mathbf{X} + \mathbf{E}_\mathbf{X} = \mathbf{X}\mathbf{W}\mathbf{P}_\mathbf{X} + \mathbf{E}_\mathbf{X}, \tag{1}$$

where $\mathbf{T}$ is an $N \times R$ component score matrix that contains the scores of the N observations on the R components, $\mathbf{P}_\mathbf{X}$ is the $R \times J$ loading matrix that contains the loadings of the predictor variables on the components, $\mathbf{E}_\mathbf{X}$ are the residual $\mathbf{X}$ scores and $\mathbf{W}$ is a $J \times R$ weight matrix. The criteria in $\mathbf{Y}$ are regressed on the components instead of on the predictors:

$$\mathbf{Y} = \mathbf{T}\mathbf{P}_\mathbf{Y} + \mathbf{E}_\mathbf{Y}, \tag{2}$$

where the matrix $\mathbf{P}_\mathbf{Y}$ ($R \times K$) contains the resulting regression weights for each of the K criteria (Kiers and Smilde 2007) and $\mathbf{E}_\mathbf{Y}$ contains the residual $\mathbf{Y}$ scores. Note that when R equals J, PCovR boils down to standard multivariate multiple regression.

To partly identify the solution (without loss of fit), the variances of the component scores (i.e., the columns of $\mathbf{T}$) are fixed at 1. This implies that in case the predictors are standardized and the components are orthogonal, the loadings in $\mathbf{P}_\mathbf{X}$ equal the correlations between the respective components and variables.

Each PCovR solution has rotational freedom. Indeed, premultiplying $\mathbf{P}_\mathbf{X}$ and $\mathbf{P}_\mathbf{Y}$ by a random transformation matrix $\mathbf{B}$ and postmultiplying $\mathbf{T}$ by $\mathbf{B}^{-1}$, does not alter the reconstructed $\mathbf{X}$ scores or the predicted $\mathbf{Y}$ scores. In empirical practice, researchers may take advantage of this rotational freedom to enhance the interpretability of the components. Specifically, they may orthogonally or obliquely rotate the loading matrix towards an a priori specified target structure (Browne 1972a,b) or towards simple structure.

A simple structure implies that there is only one non-zero loading per variable and there are more than R and fewer than J zero-loadings per component (Browne 2001). The components can then be labeled by considering what the variables with a clear non-zero loading on a component have in common. To approximate simple structure, several rotation criteria have been proposed. For the PCovR case the following criteria seem useful: Varimax, Weighted Varimax, Quartimin, Simplimax, and Promin.

- Varimax (Kaiser 1958) is a very popular orthogonal rotation criterion that maximizes the sum of the variances of the squared loadings:

$$f(\mathbf{P_X}) = \sum_{r=1}^{R} \left[\frac{1}{J} \sum_{j=1}^{J} \left(p_{rj}^2 - \overline{p_r^2} \right)^2 \right]. \tag{3}$$

- Weighted Varimax (Cureton and Mulaik 1975) is an oblique variant of Varimax, in which the simplest variables (i.e., the ones with only one high loading) have more influence on the rotation than the complex variables (i.e., the ones with multiple high loadings).

- Quartimin (Carroll 1953) is an oblique rotation strategy that minimizes the sum across variables and across component pairs of the cross-products of the squared loadings:

$$f(\mathbf{P_X}) = \sum_{j=1}^{J} \sum_{r_1=1}^{R} \sum_{r_2 \neq r_1}^{R} p_{jr_1}^2 p_{jr_2}^2. \tag{4}$$

- Simplimax (Kiers 1994) finds the target matrix that can be approximated best by rotation, among all simple-structure target matrices (i.e., matrices with as few as possible non-zero loadings per variable) that have a specified number of zero elements. This number can be chosen a priori or by comparing the rotation function values for different numbers of zeros, retaining the number after which the improvement in function value levels off (similarly to the scree test procedure).

- Promin (Lorenzo-Seva 1999) applies oblique target rotation using the Weighted Varimax solution as the target matrix.

2.3. Loss function

One of the key features of PCovR is that the reduction of the predictors to components and the prediction of the criteria by those components is conducted simultaneously. To this end, a weighting parameter α is used, ranging between 0 and 1, that determines to what degree the reduction and prediction parts of the model are emphasized. Specifically, in a PCovR analysis, the following loss function value L is minimized:

$$L = \alpha \frac{\|\mathbf{X} - \mathbf{TP_X}\|^2}{\|\mathbf{X}\|^2} + (1 - \alpha) \frac{\|\mathbf{Y} - \mathbf{TP_Y}\|^2}{\|\mathbf{Y}\|^2}, \tag{5}$$

with $\|\mathbf{A}\|$ being the Frobenius matrix norm of a matrix $\mathbf{A}$. Note that PCovR analyses with α values of 0 and 1 correspond, respectively, to reduced-rank regression and principal component regression (Smilde and Kiers 1999).

2.4. Estimation

Given a specific α value and number of components R, a closed form solution exists for estimating the PCovR parameters. Specifically, $\mathbf{T}$ is estimated by computing the first R eigenvectors of the matrix

$$\mathbf{G} = \alpha \frac{\mathbf{X}\mathbf{X}^\top}{\|\mathbf{X}\|^2} + (1-\alpha)\frac{\mathbf{H}_\mathbf{X}\mathbf{Y}\mathbf{Y}^\top\mathbf{H}_\mathbf{X}}{\|\mathbf{Y}\|^2}, \tag{6}$$

in which $\mathbf{H}_\mathbf{X}$ is the projection matrix which projects $\mathbf{Y}$ on $\mathbf{X}$. $\mathbf{P}_\mathbf{X}$ and $\mathbf{P}_\mathbf{Y}$ can then be calculated, respectively, as

$$\mathbf{P}_\mathbf{X} = \mathbf{T}^\top\mathbf{X} \tag{7}$$

and

$$\mathbf{P}_\mathbf{Y} = \mathbf{T}^\top\mathbf{Y} \tag{8}$$

(De Jong and Kiers 1992). Finally, $\mathbf{T}$, $\mathbf{P}_\mathbf{X}$, and $\mathbf{P}_\mathbf{Y}$, are rescaled such that the columns of $\mathbf{T}$ have a variance of 1.

2.5. Model selection

As the most appropriate number of components R is often unknown in practice, PCovR model selection involves two decisions: selecting the number of components and tuning the α value. In the literature, most authors solve this procedure by performing cross-validation (Smilde and Kiers 1999; De Jong and Kiers 1992). As this simultaneous selection procedure may become rather time-consuming, in case the data are large and one considers many different R and α values, we also propose a new and fast sequential procedure, in which we first select the α value using maximum likelihood principles and, subsequently, the number of components by means of a generalization of the well-known scree test. Of course, also substantive arguments based on interpretability of the obtained solution, previous research or theory can be taken into account (Ceulemans and Kiers 2006).

Simultaneous procedure

The simultaneous procedure selects the optimal α and R values by performing leave-one-out cross-validation (Hastie, Tibshirani, and Friedman 2001). In leave-one-out cross-validation, one conducts N PCovR analyses, each time discarding a different observation. Next, for each discarded observation, reconstructed $\mathbf{y}_n^{CV}$ scores are computed given the PCovR estimates for the other observations:

$$\mathbf{y}_n^{CV} = \mathbf{x}_n\mathbf{W}_n\mathbf{P}_{\mathbf{y},n}, \tag{9}$$

where $\mathbf{x}_n$ contains the predictor scores of the nth observation, and $\mathbf{W}_n$ and $\mathbf{P}_{\mathbf{Y},n}$ indicate the $\mathbf{W}$ and $\mathbf{P}_\mathbf{Y}$ matrix of the analysis in which the nth observation was discarded. Finally, the leave-one-out cross-validation fit is calculated as

$$Q^2_\mathbf{Y} = 1 - \frac{\sum_{n=1}^N \|\mathbf{y}_n - \mathbf{y}_n^{CV}\|^2}{\|\mathbf{Y}\|^2}. \tag{10}$$

This is done for each combination of an α and an R value, and the α and R values that maximize $Q^2_\mathbf{Y}$ are retained. Note that this strategy requires a relatively high computational effort, because the PCovR analysis needs to be performed N times for each (α,R) combination

that is considered. In order to save computation time, it is possible to perform k-fold cross-validation (Hastie *et al.* 2001). This implies that one discards more than one observation in each of the k cross-validation steps by splitting the data in k roughly equal-sized parts and omitting all observations that belong to a particular part in the corresponding step.

Sequential procedure

In the sequential selection procedure, we first tune α on the basis of maximum likelihood principles (Vervloet, Van Deun, Van den Noortgate, and Ceulemans 2013): Given the assumption that the error on the predictor block ($\mathbf{E_X}$) and the error on the criterion block ($\mathbf{E_Y}$) is drawn from a normal distribution with mean 0 and variances of, respectively, $\sigma^2_{\mathbf{E_X}}$ and $\sigma^2_{\mathbf{E_Y}}$, the α value that will maximize the likelihood of the data given the model is equal to:

$$\alpha_{ML} = \frac{\|\mathbf{X}\|^2}{\|\mathbf{X}\|^2 + \|\mathbf{Y}\|^2 \frac{\sigma^2_{\mathbf{E_X}}}{\sigma^2_{\mathbf{E_Y}}}}. \tag{11}$$

This value is approximated by replacing the variances $\sigma^2_{\mathbf{E_X}}$ and $\sigma^2_{\mathbf{E_Y}}$ by estimates. An estimate for $\sigma^2_{\mathbf{E_X}}$ can be obtained by applying principal component analysis to $\mathbf{X}$ and determining the optimal number of components through a scree test (see Section 3); the estimate equals the associated percentage of unexplained variance. The estimate for $\sigma^2_{\mathbf{E_Y}}$ boils down to the percentages of unexplained variance when $\mathbf{Y}$ is regressed on $\mathbf{X}$. This approach for estimating $\sigma^2_{\mathbf{E_X}}$ and $\sigma^2_{\mathbf{E_Y}}$ was based on the work of Wilderjans, Ceulemans, Van Mechelen, and Van den Berg (2011).

Next, to select the optimal (i.e., good fit without being overly complex) number of components, we compute the solutions with $R_{\min} - 1$ to $R_{\max} + 1$ components and the corresponding weighted sum of the percentage of variance accounted for in $\mathbf{X}$ and in $\mathbf{Y}$:

$$VAF_{R,sum} = \alpha \frac{\|\mathbf{T}^R {\mathbf{P_X}}^R\|^2}{\|\mathbf{X}\|^2} + (1-\alpha) \frac{\|\mathbf{T}^R {\mathbf{P_Y}}^R\|^2}{\|\mathbf{Y}\|^2}. \tag{12}$$

Note that for $R = 0$, this sum equals 0. Subsequentially, among the $R_{\min}$ to $R_{\max}$ values, the optimal R value is the one that has the highest st value (Cattell 1966; Wilderjans, Ceulemans, and Meers 2013):

$$st_R = \frac{VAF_{R,sum} - VAF_{R-1,sum}}{VAF_{R+1,sum} - VAF_{R,sum}}. \tag{13}$$

Indeed, when plotting $VAF_{R,sum}$ as a function of R, the solution with the highest st value will be the one after which $VAF_{R,sum}$ levels off, implying that additional components do not significantly increase the fit of the solution.

The sequential procedure can be amended with cross-validation steps. First, when selecting the optimal R value given α_{ML}, one can replace the scree ratio step by a step in which the cross-validation fit for different R values is computed and the R value that maximizes this fit is retained. Alternatively, once the optimal R value given α_{ML}, has been determined in the scree step, one can add a third step that consists of a cross-validation procedure to empirically assess the α value that optimizes the prediction of future data. Indeed, α_{ML} may be inaccurate if the assumptions about the error variances – each predictor (resp., criterion) has the same error variance, the error of different predictors (resp., criteria) is uncorrelated – are violated.

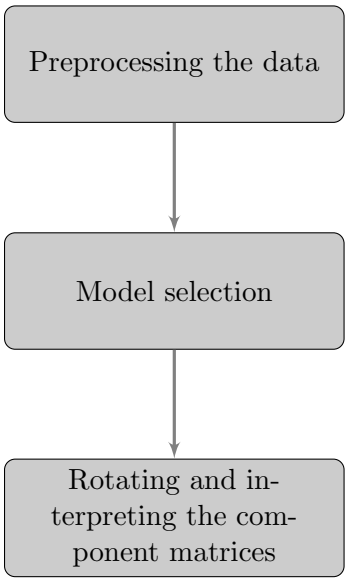

Figure 1: Flowchart of the different steps in a PCovR analysis.

In practice, it may happen that these four procedures point towards a different solution, as will also be the case for our illustrative dataset (see below). Indeed, the scree test based procedures favor a solution with a few components that each explain a lot of variance whereas the cross-validation based methods focus on the prediction of future data, which can of course (slightly) improve by including components that explain a small amount of variability only, but in a consistent way. In such cases, following Hastie *et al.* (2001), we recommend to retain the more parsimonious model among the indicated models (i.e., fewer components) if the corresponding cross-validation fit is only slighly lower.

3. Using the PCovR package

In this section, we discuss how a PCovR analysis can be conducted in R. First, one loads the **PCovR** package:

```
R> library("PCovR")
```

To load the dataset `psychiatrists` and start the analysis, one types

```
R> data("psychiatrists", package = "PCovR")
R> X <- psychiatrists$X
R> Y <- psychiatrists$Y
R> results <- pcovr(X, Y)
```

which runs the analysis with the default analysis settings and saves the output in a list variable, that is called `results` here. This command requires $\mathbf{X}$ and $\mathbf{Y}$ to be numerical data frames. Of course, as will be explained below, the command can be further refined, to use other analysis settings.

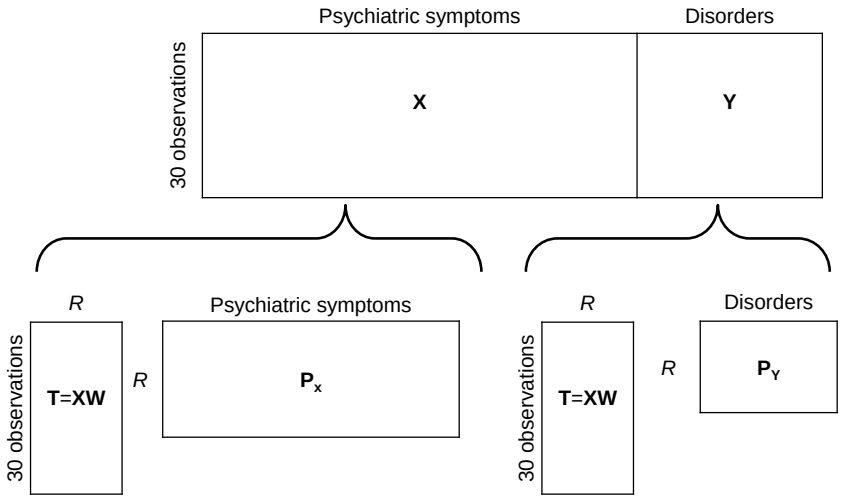

Figure 2: Visualization of the real-data example and the PCovR decomposition.

The three main steps of a PCovR analysis are: (1) preprocessing the data, (2) determining the number of components R and choosing the α weight, and (3) rotating and interpreting the solution (Figure 1). These steps will be illustrated with a real-data example, called `psychiatrists` (Van Mechelen and De Boeck 1990), that is included in the **PCovR** package. The data contain the scores of 30 Belgian psychiatric patients on 23 psychiatric symptoms and 4 psychiatric disorders (toxicomania, schizophrenia, depression, and anxiety disorder). Each score is the summated score of the binary symptom and disorder scores that were given by 15 different psychiatrists. In our analysis, we will examine the extent to which the degree of toxicomania, schizophrenia, depression and anxiety disorder, can be predicted by the 23 psychiatric symptoms (see Figure 2). Applying ordinary least squares regression (OLS) would not be appropriate, because the different symptoms are very likely to have a high degree of multicollinearity (indeed, 18 of the 23 variance inflation factors are larger than 5). Furthermore, PCovR is useful here to gain insight into the correlation structure of the data.

3.1. Preprocessing the data

The **PCovR** package includes two preprocessing options, which can be applied to $\mathbf{X}$ and/or $\mathbf{Y}$. Specifically, it is possible to only center the data (`prepX = "cent"`, `prepY = "cent"`). However, the default option is to standardize the data (`prepX = "stand"`, `prepY = "stand"`), which implies that $\mathbf{X}$ and/or $\mathbf{Y}$ are centered and normalized (i.e., each variable has a mean of zero and a standard deviation of one).

3.2. Model selection

The package offers the simultaneous and sequential model selection procedures that were described in the Section 2. A sequential approach needs less computational time as is shown in Table 1 for the `psychiatrists` dataset. Note that the assessment of the optimal R value can be overruled, in case one is only interested in the solutions with a particular R value. In particular, when specifying the input parameter `R`, `Rmin` and `Rmax` will be ignored, and the specified number of components will be used when running the analysis and determining α.

	Selected α	Selected R	$Q^2_{\mathbf{Y}}$	Computation time
`modsel = "sim"`	0.79	6	0.73	117.47 s
`modsel = "seqAcv"`	0.49	3	0.68	16.92 s
`modsel = "seqRcv"`	0.21	4	0.64	1.22 s
`modsel = "seq"`	0.21	3	0.63	0.16 s

Table 1: Overview of the available model selection procedures and their outcomes for the `psychiatrists` dataset. The analyses were run with R version 3.0.2 using an Intel Core 2 Duo P9700 processor.

Simultaneous procedure

The simultaneous procedure (`modsel = "sim"`) that was explained earlier performs leave-one-out cross-validation for a range of α and R values. By default, 100 α values between 0.01 and 1 are explored, but alternatively, a vector (or scalar) of choice can be specified with the parameter `weight`. The same holds for the number of components, which is by default initialized as $R_{\text{min}} = 1$ and $R_{\text{max}} = J/3$. The α and R value combination that maximizes $Q^2_{\mathbf{Y}}$ is retained. Note that the parameter `fold` can be used to alter the number of roughly equal-sized parts in which the data are split for cross-validation. The default value of `fold` is `"LeaveOneOut"`, implying that the data is split in N parts. For the `psychiatrists` dataset, the 6-component solution with $\alpha = 0.79$ has the highest cross-validation fit.

Sequential procedure

The fastest and therefore default model selection setting (`modsel = "seq"`) implies a sequential procedure in which α is determined on the basis of maximum likelihood principles (11), unless a specific α value is imposed by the user (e.g., `weight = 0.50`). For instance, for the `psychiatrists` dataset, $\alpha_{ML} = 0.21$ if the error variance ratio is determined on the basis of a principal component analysis of $\mathbf{X}$ and a regression of $\mathbf{Y}$ on $\mathbf{X}$ (see Section 2), which is the default option. Among all models with the selected α value and a number of components between R_{min} and R_{max}, the solution is picked that has the highest *st* value (13), which is the 3-component solution when using the default R_{min} and R_{max} values.

The package also provides two sequential procedures that incorporate a cross-validation step (`modsel = "seqRcv"` and `modsel = "seqAcv"`). `seqRcv` also starts with the selection of α based on maximum likelihood principles, but in the next step, R is determined using leave-one-out cross-validation. `"seqAcv"` is identical to the default procedure, but has an extra step: after the selection of R using α_{ML}, leave-one-out cross-validation is applied to choose the α value among the values specified in `weight`. For the `psychiatrists` example, these procedures retain a solution with $\alpha = 0.21$ and $R = 4$ and with $\alpha = 0.49$ and $R = 3$, respectively.

For this particular dataset, it can be seen that the simultaneous selection procedure points towards more components than the three sequential procedures. To better understand this difference, it is instructive to inspect Figure 3 which displays the cross-validation fit for all models:

```
R> plot(pcovr(X, Y, modsel = "sim"))
```

This plot indeed shows that the best cross-validation fit is achieved with a 6-component

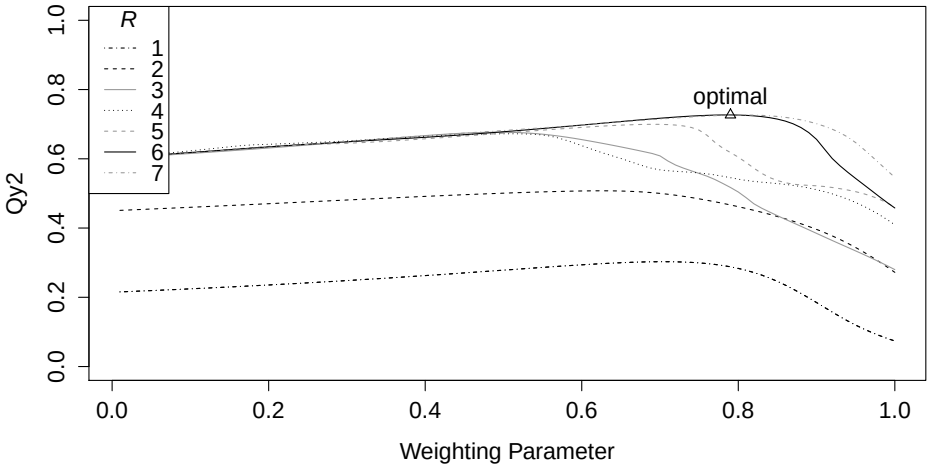

Figure 3: Cross-validation fit for the results of the simultaneous model selection procedure.

model and a relatively high α value. However, the cross-validation fit of the solutions that are retained by the sequential procedures are only slightly lower, whereas these solutions are clearly more parsimonious. Therefore, following Hastie *et al.* (2001) who recommend to also have a look at models with similar cross-validation fits, but lower complexity, we decided to retain the 3-component model with α equal to 0.49. Indeed, among the solutions retained by the sequential approaches, this solution has the highest cross-validation fit.

3.3. Interpreting the component matrices

The **PCovR** package includes seven different rotation strategies: Varimax (`rot = "varimax"`), which is the default option, Weighted Varimax (`rot = "wvarim"`), orthogonal target rotation (`rot = "TargetT"`), oblique target rotation (`rot = "TargetQ"`), Quartimin (`rot = "Quartimin"`), Simplimax (`rot = "simplimax"`), and Promin (`rot = "promin"`); one can also request the unrotated solution by typing `rot = "none"`.

Some of the rotation criteria require extra input arguments. For both orthogonal and oblique target rotation, a target matrix has to be specified with the argument `target`. When using Simplimax, the user has to specify a number of zero elements with the argument `zeroloads` (which equals J by default).

The interpretation of a specific PCovR solution usually starts with the inspection of the loading matrix, which can be requested by the command

```
R> results$Px
```

For instance, Table 2, which displays the Varimax rotated loadings of the $\alpha = 0.49$ and $R = 3$ solution for the `psychiatrists` data, reveals that the first component has highly positive loadings for all depressive symptoms (e.g., "depression", "suicide" and "social isolation") and a negative loading for hallucinations. The second component seems to indicate the amount of substance abuse (e.g., "narcotics" and "alcohol"), while the third component reflects inappropriate behavior (e.g., "inappropriate", "social leveling", "desorganised speech", "routine") versus fear. These three components are uncorrelated. Note that Varimax was used here, be-

	1st component	2nd component	3rd component
depression	**0.94**	−0.10	0.13
suicide	**0.68**	−0.02	0.09
hallucinations	**−0.46**	−0.30	−0.29
social_isolation	**0.42**	−0.28	**−0.42**
grandeur	**−0.39**	0.29	0.10
antisocial	**−0.37**	**0.36**	0.09
occupational_dysfunction	0.34	−0.11	−0.32
negativism	−0.29	−0.07	−0.16
desorientation	0.21	−0.03	−0.19
somatic_concern	0.19	−0.16	0.00
narcotics	−0.01	**0.91**	0.00
alcohol	−0.22	**0.67**	−0.01
suspicion	−0.06	−0.26	−0.18
agitation	−0.02	−0.18	−0.05
fear	0.23	−0.32	**0.81**
social_leveling	0.12	−0.33	**−0.61**
desorganized_speech	−0.26	−0.22	**−0.54**
denial	**−0.39**	−0.11	**−0.53**
inappropriate	−0.29	−0.25	**−0.51**
retardation	0.23	−0.24	**−0.45**
routine	0.14	−0.32	**−0.42**
intellectual_obstruction	0.14	0.00	−0.32
impulse_control	−0.07	−0.11	−0.13
toxicomania	−0.08	**0.98**	−0.04
schizophrenia	**−0.52**	**−0.46**	**−0.64**
depression	**0.97**	−0.04	0.08
anxiety_disorder	0.24	−0.07	**0.90**

Table 2: Varimax rotated loadings and associated regression weights of the $\alpha = 0.49$ and $R = 3$ solution for the `psychiatrists` dataset. The highest loadings (i.e., with an absolute value higher than 0.35) are shown in bold.

cause it is the only built-in exploratory rotation criterion that yields orthogonal components, which enhances the interpretability of the regression coefficients, and because the resulting component loadings were sufficiently clear.

After labeling the components, the regression weights (`results$Py`) and component scores (`results$Te`) can also be interpreted. The regression weights indicate to which extent the criteria can be predicted on the basis of the components and the component scores reflect how each individual scores on these components. From the regression weights for the `psychiatrists` dataset in Table 2, it can be concluded that the degree of depressive symptomatology versus hallucinations of individuals is a strong predictor of both depression (positive relation) and schizophrenia (negative relation). Substance abuse can predict both toxicomania (positive relation) and schizophrenia (negative relation). The third component, inappropriate behavior versus fear is associated with schizophrenia (positive relation) and anxiety disorder (negative relation).

4. Conclusion

The main features of the R package **PCovR** have been explained and illustrated in this paper, using the dataset `psychiatrists` that is available in the package. **PCovR** is a package for performing principal covariates regression, a method developed by De Jong and Kiers (1992). The package depends on the packages **GPArotation** (Bernaards and Jennrich 2005), **ThreeWay** (Giordani, Kiers, and Del Ferraro 2014), **MASS** (Venables and Ripley 2002), and **Matrix** (Bates and Mächler 2014).

Acknowledgments

The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003), and by the Interuniversity Attraction Poles program financed by the Belgian government (IAP/P7/06).

References

Asparouhov T, Muthén B (2009). "Exploratory Structural Equation Modeling." *Structural Equation Modeling*, **16**(3), 397–438.

Bates D, Mächler M (2014). ***Matrix**: Sparse and Dense Matrix Classes and Methods*. R package version 1.2-0, URL http://CRAN.R-project.org/package=Matrix.

Bernaards CA, Jennrich RI (2005). "Gradient Projection Algorithms and Software for Arbitrary Rotation Criteria in Factor Analysis." *Educational and Psychological Measurement*, **65**, 676–696.

Browne MW (1972a). "Orthogonal Rotation to a Partially Specified Target." *British Journal of Mathematical and Statistical Psychology*, **25**(1), 115–120.

Browne MW (1972b). "Oblique Rotation to a Partially Specified Target." *British Journal of Mathematical and Statistical Psychology*, **25**(2), 207–212.

Browne MW (2001). "An Overview of Analytic Rotation in Exploratory Factor Analysis." *Multivariate Behavioral Research*, **36**(1), 111–150.

Carroll JB (1953). "An Analytical Solution for Approximating Simple Structure in Factor Analysis." *Psychometrika*, **18**(1), 23–38.

Cattell RB (1966). "The Scree Test for the Number of Factors." *Multivariate Behavioral Research*, **1**(2), 245–276.

Ceulemans E, Kiers HAL (2006). "Selecting among Three-Mode Principal Component Models of Different Types and Complexities: A Numerical Convex Hull Based Method." *British Journal of Mathematical and Statistical Psychology*, **59**(1), 133–150.

Cureton EE, Mulaik SA (1975). "The Weighted Varimax Rotation and the Promax Rotation." *Psychometrika*, **40**(2), 183–195.

De Jong S, Kiers HAL (1992). "Principal Covariates Regression: Part I. Theory." *Chemometrics and Intelligent Laboratory Systems*, **14**(1–3), 155–164.

Giordani P, Kiers HAL, Del Ferraro MA (2014). "Three-Way Component Analysis Using the R Package **ThreeWay**." *Journal of Statistical Software*, **57**(7), 1–23. URL http://www.jstatsoft.org/v57/i07/.

Hastie T, Tibshirani R, Friedman J (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York.

Izenman AJ (1975). "Reduced-Rank Regression for the Multivariate Linear Model." *Journal of Multivariate Analysis*, **5**(2), 248–264.

Jolliffe IT (1982). "A Note on the Use of Principal Components in Regression." *Journal of the Royal Statistical Society C*, **31**(3), 300–303.

Kaiser HF (1958). "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika*, **23**(3), 187–200.

Kiers HAL (1994). "Simplimax: Oblique Rotation to an Optimal Target with Simple Structure." *Psychometrika*, **59**(4), 567–579.

Kiers HAL, Smilde AK (2007). "A Comparison of Various Methods for Multivariate Regression with Highly Collinear Variables." *Statistical Methods and Applications*, **16**(2), 193–228.

Lorenzo-Seva U (1999). "Promin: A Method for Oblique Factor Rotation." *Multivariate Behavioral Research*, **34**(3), 347–365.

Mevik BH, Wehrens R, Liland KH (2013). ***pls**: Partial Least Squares and Principal Component Regression*. R package version 2.4-3, URL http://CRAN.R-project.org/package=pls.

Muthén LK, Muthén BO (1998–2011). ***Mplus** User's Guide*. 6th edition. Muthén & Muthén, Los Angeles.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

Smilde AK, Kiers HAL (1999). "Multiway Covariates Regression Models." *Journal of Chemometrics*, **13**(1), 31–48.

Van Mechelen I, De Boeck P (1990). "Projection of a Binary Criterion into a Model of Hierarchical Classes." *Psychometrika*, **55**(4), 677–694.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S*. 4th edition. Springer-Verlag, New York. URL http://www.stats.ox.ac.uk/pub/MASS4/.

Vervloet M, Kiers HAL, Ceulemans E (2015). ***PCovR**: Principal Covariates Regression*. R package version 2.6, URL http://CRAN.R-project.org/package=PCovR.

Vervloet M, Van Deun K, Van den Noortgate W, Ceulemans E (2013). "On The Selection of the Weighting Parameter Value in Principal Covariates Regression." *Chemometrics and Intelligent Laboratory Systems*, **123**, 36–43.

Wilderjans TF, Ceulemans E, Meers K (2013). "CHull: A Generic Convex-Hull-Based Model Selection Method." *Behavior Research Methods*, **45**(1), 1–15.

Wilderjans TF, Ceulemans E, Van Mechelen I, Van den Berg RA (2011). "Simultaneous Analysis of Coupled Data Matrices Subject to Different Amounts of Noise." *British Journal of Mathematical and Statistical Psychology*, **64**(2), 277–290.

Wold S, Ruhe A, Wold H, Dunn III WJ (1984). "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses." *SIAM Journal of Statistics and Computations*, **5**(3), 735–743.

Affiliation:

Marlies Vervloet
Methodology of Educational Sciences Research Group
KU Leuven
Tiensestraat 102 bus 3762
3000 Leuven, Belgium
E-mail: marlies.vervloet@ppw.kuleuven.be

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 65, Issue 8 *Submitted:* 2013-03-22
May 2015 *Accepted:* 2014-09-17